This article has been accepted for publication in IEEE Internet of Things Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2025.3535732

Mixed Vehicle Platoon Forming: A Multi-agent Reinforcement Learning Approach

Yujie Shi, Haoxuan Dong, *Member, IEEE*, Chaozhe R. He, *Member, IEEE*, Yuxiao Chen, *Member, IEEE*, Ziyou Song, *Senior Member, IEEE*

***Abstract*—The emerging connected and automated vehicle (CAV) technologies present opportunities to improve traffic safety, economy, and efficiency. However, the diverse uncontrollable human-driven vehicles (HDVs) will continue to predominate traffic for a long time, resulting in coexisting CAVs and HDVs in the form of mixed vehicle platoons. This study proposes a mixed vehicle platoon forming method based on a two-stage control framework to adapt to dynamic mixed traffic environments. The platoon formation generation stage creates feasible formation (i.e., a spatially coordinated mix of CAVs and HDVs ensuring safe and efficient platoon control.) appropriate for mixed traffic based on the empirical formation method. The multi-agent reinforcement learning is used in the second stage for realizing the platoon forming control safely and efficiently with the guidance of feasible formation. Finally, extensive simulations are conducted using the Highway-env simulator to evaluate the effectiveness of the proposed method. The results demonstrate that the proposed method can effectively control CAVs in conjunction with HDVs to form a mixed vehicle platoon, achieving an average energy efficiency improvement of up to 10.69% and a reduction in travel time by up to 2.73% compared to benchmark strategies.**

***Index Terms*—Connected and automated vehicle, mixed traffic, reinforcement learning, platoon forming, energy-efficient driving.**

I. Introduction

The emerging connected and automated vehicle (CAV) technologies present opportunities to significantly improve traffic safety [1], economy [2], and efficiency [3]. However, the diverse uncontrollable human-driven vehicles (HDVs) will continue to predominate traffic for a considerable amount of time, resulting in mixed CAVs and HDVs presence in the form of mixed vehicle platoons [4]. In contrast to the pure CAV platoon [5, 6], the efficient control of a mixed vehicle platoon is extremely difficult due to the highly random and unpredictable driving behaviors of HDVs [7, 8].

By regulating the CAVs, mixed vehicle platoon control seeks to maximize vehicle driving safety, energy efficiency, and traffic efficiency. To accomplish these objectives, Mousavi *et al*. [9] proposed a gain-scheduled control strategy that guides the acceleration of a single CAV to improve traffic flow stability, i.e., mitigating traffic disruptions and coping with parametric uncertainties. Xue *et al*. [10] implemented a distributed optimal cooperative control algorithm for multiple CAVs to enhance traffic efficiency in mixed traffic. However, these studies mentioned above only focus on the optimization of CAVs, while neglecting the impact of HDV driving behaviors. As a result, the efficient driving control potential of mixed vehicle platoon is largely unexplored and is not even an optimal solution for the platoon as a whole. To this end, Wang *et al*. [11] and Li *et al*. [12] presented adaptive cruise control for mixed traffic strategies with the consideration of the impact of CAV's efficient driving on the following HDVs, the unified optimization framework with reinforcement learning (RL) is formulated for improving the holistic energy efficiency of the mixed vehicle platoon. Li *et al*. [13] reduced collision risks near freeway bottlenecks for mixed traffic in two-lane scenarios, which is realized by the dynamic speed control of all CAVs with the consideration of HDVs. In addition, recent research has focused on developing methods to model and control the propagation of disturbances and ensure safety within mixed vehicle platoons [14, 15]. These efforts aim to address uncertainties introduced by HDVs while optimizing coordination among all vehicles in the platoon.

Despite the valuable studies of mixed traffic regulation that have been provided, they predominantly concentrate on specific platoon formations, which are the fixed spatial distribution of CAVs and HDVs in the platoon. For example, HDV–CAV–HDV [11], HDV–CAV–HDV–CAV–HDV [12], and CAV–HDV–HDV–… [16]. The study done by Zheng *et al*. [17] and Yang *et al*. [18] demonstrated that vehicle driving safety, energy efficiency, and traffic efficiency of mixed traffic flow are affected by the spatial distribution of CAVs. In this context, a hindrance-aware platoon formation method for mixed traffic was presented by Zhu *et al.* [19]. It aimed to coordinate CAVs to maximize overall traffic flow while lowering collision risk and fuel usage. Jin *et al*. [20] explored the impact of three formations on energy and traffic efficiencies in mixed traffic. Li *et al.* [21] investigated the platoon formation of CAVs impact traffic performance from the standpoint of set-function optimization, which considered uniform distribution formation, random formation, and platoon formation. Jin *et al*. [20] and Li *et al*. [21] both indicate that the spatial distribution of platoons in mixed traffic flow has a significant impact on overall energy consumption and traffic efficiency.

The spatial distribution of HDVs and CAVs is random in real-world traffic. Even though the aforementioned research has produced insightful studies on mixed vehicle platoons, they

Manuscript received August 2024. This work is funded by the A-Star Young Individual Research Grants (YIRG), Singapore, under Grant M22K3c0092. (Yujie Shi and Haoxuan Dong are co-first authors.) (*Corresponding author: Ziyou Song.*)

Yujie Shi, Haoxuan Dong, and Ziyou Song are with the Department of Mechanical Engineering, National University of Singapore, Singapore 117575, Singapore (e-mail: yujie.shi@u.nus.edu; donghaox@foxmail.com; ziyou@nus.edu.sg).

Chaozhe R. He is with the Department of Mechanical and Aerospace Engineering, University at Buffalo, State University of New York, Buffalo, NY 14260 USA (e-mail: chaozheh@buffalo.edu).

Yuxiao Chen is with NVIDIA Research, NVIDIA Corporation, Santa Clara, CA 95051 USA (e-mail: yuxiaoc@nvidia.com).

have not taken into consideration the forming process by which platoons form from a disordered distribution of HDVs and CAVs to a designated distribution. Practical traffic flow typically has time-varying states and the HDV drives dynamically due to stochastic lane-changing and lane-keeping behaviors [22], which influences the CAV motion trajectory control. This poses the challenge of how to control the motion of CAVs to form the designated formation of mixed traffic. For this problem, Woo *et al*. [23] and Cai *et al.* [24] designed a platoon organization strategy for mixed traffic, which can form a pure CAV platoon in designated lanes from randomly distributed CAVs in multiple lanes. However, the method of specifying CAVs to form a pure CAV platoon formation may not be able to achieve the holistic optimal of mixed traffic. Few studies have been conducted on mixed vehicle platoon forming, Wu *et al.* [25] designed a cooperative strategy to harmonize the spatial distribution of CAVs in mixed traffic, to improve the crossing efficiency of unsignalized intersections. Maiti *et al.* [26] developed Ad-hoc platoon formation and dissolution strategies to encourage heavy-duty vehicles to drive in platoons for multi-lane highways, which can reduce fuel usage. The rule-based mixed vehicle platoon forming method proposed by Wu *et al.* [25] and Maiti *et al.* [26] is computationally efficient; however, it cannot adapt to the dynamic mixed traffic due to the irregular spatial distribution in HDVs and CAVs as well as the random driving behaviors of HDVs.

This study proposes a mixed vehicle platoon forming method using multi-agent reinforcement learning (MARL) that considers both the dynamic driving behaviors of HDVs and designated spatial distribution. The major contributions are threefold. First, a hierarchical control framework is proposed for achieving mixed vehicle platoon forming control in dynamic mixed traffic, which includes two stages, i.e., mixed vehicle platoon formation generation and forming control. Second, considering CAVs as agents, an efficient mixed vehicle platoon forming method is proposed using MARL to address the challenging CAV regulation problem in the presence of random HDV perturbations. The third contribution is a comprehensive evaluation conducted on various perspectives of mixed vehicle platoon, including forming success, driving safety, and energy and traffic efficiencies. The dynamic mixed traffic is simulated using the Highway-env simulator [27].

The rest of this paper is organized as follows. Section II provides a description of the traffic scenario, mixed platoon model, and RL theory. Section III introduces the research problem and the control framework. In Section IV, the methodology of the mixed vehicle platoon forming control is formulated. The performance of the proposed method is evaluated by simulations in Section V. Finally, Section VI concludes this study.

II. Problem Statement and Control Framework

A. Problem Statement

Some studies have provided optimal mixed vehicle platoon formations that can improve traffic safety, economy, and efficiency [20]. However, forming the designated platoon formation from a disordered spatial distribution of CAVs and HDVs poses two challenges in real-world traffic. This is a result of the irregular spatial distribution in HDVs and CAVs as well as the random driving behaviors of HDVs.

1) Challenge 1: How to determine the order in which CAVs form the platoon? The initial position of CAVs in traffic is random, and as Fig. 1(a) illustrates, CAVs in Lane 1 and Lane 3 can merge into Lane 2 to form the mixed vehicle platoon. The spatial distribution of CAVs in a mixed vehicle platoon varies according to the platoon formation; selecting which CAV to merge into a given position is an extremely challenging decision-making task that impacts traffic safety and platoon forming effectiveness.

2) Challenge 2: What is the suitable CAV driving trajectory for platoon forming? As shown in Fig. 1(b), the CAV in the Lane 3 is controlled to merge into the Lane 2 to form a mixed vehicle platoon. Three different trajectories are possible for the CAV lane-changing trajectory, denoted as No.1, No.2, and No. 3. However, due to the random driving behavior of HDVs, only the No.2 trajectory is practical because of the risk of collision from surrounding HDVs in Lane 2 and Lane 3. Planning a suitable driving trajectory is difficult since CAVs are formed into platoons with a range of driving trajectories. These trajectories differ in terms of traffic safety and efficiency, which are affected by the surrounding vehicles and the movement of rear vehicles in the target lane.

Therefore, the control of CAVs during platoon forming is not only determined by designated formation but also influenced by the surrounding vehicles. In addition, the CAV driving trajectory affects the following vehicle's behavior. These factors increase the risk of collision and potentially disturb traffic flow.

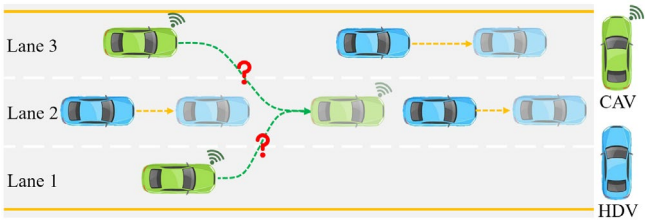
B. Control Framework

This study proposes a mixed vehicle platoon forming method using a hierarchical control framework to realize safe and efficient platoon forming control, as shown in Fig. 2. The proposed method is composed of two stages, i.e., mixed vehicle platoon formation generation and forming control. In the first stage, the feasible formation sequence is generated based on the empirical formation method and the current states of CAVs and HDVs, which is the guidance of the platoon forming control. Note that a feasible formation refers to a coordination of CAVs and HDVs, designed based on their spatial distribution to ensure safe and efficient control of the mixed vehicle platoon. MARL is used in the second stage for realizing the platoon forming control efficiently. Following the platoon formation requirements, CAVs in different lanes perform lane-changing or lane-keeping operations under the policy obtained from MARL to drive into the designated positions within the mixed vehicle platoon in dynamic traffic.

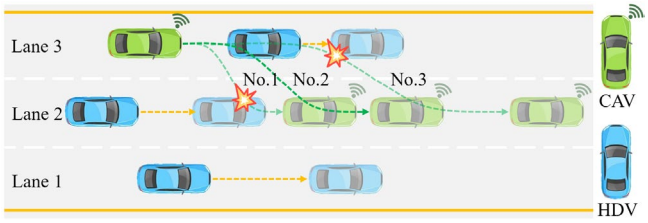
III. Preliminaries

A. Traffic Scenario

We define a generic route with three lanes that includes mixed traffic flow, as shown in Fig. 3, where the lane numbers from the outside to the inside of the road are $i = 1, 2, 3$, the width of each lane is D_l, and the maximum and minimum speed limits of the road are V_{max} and V_{min}, respectively.



(a) How to determine the order in which CAVs form the platoon?



(b) What's a suitable CAV driving trajectory for platoon forming?

Fig. 1. Schematic diagram of CAV driving for mixed vehicle platoon forming control. Here, the dark-colored vehicles indicate the state at time t and the light-colored vehicles indicate the state at time $t + \Delta t$.

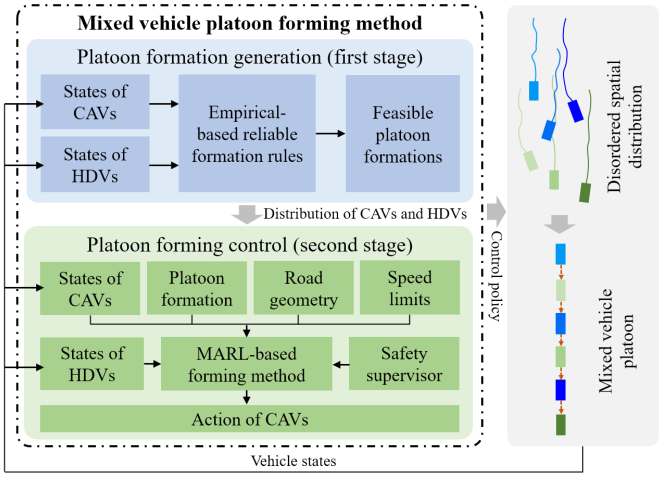


Fig. 2. The control framework of mixed vehicle platoon forming method.

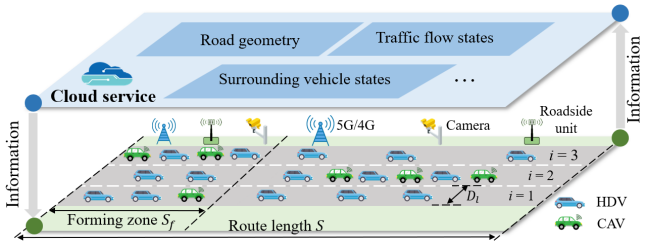


Fig. 3. Scheme of the route with mixed traffic flow.

The mixed traffic flow consists of both controllable CAVs and uncontrollable HDVs. In the environment of the Internet of Vehicles, the CAVs can be connected using vehicle-to-vehicle communication. The HDVs' states and traffic flow states can be sensed by roadside cameras and uploaded to the Cloud. Additionally, high-precision maps stored in the cloud can be utilized to acquire road geometry data [28]. Leveraging the Internet of Vehicles technology, essential data for mixed vehicle forming control, such as road geometry, traffic flow states, and surrounding vehicle states, can be gathered. Based on this information, control commands are generated in the cloud, enabling the efficient coordination of CAVs.

The route can be divided into two zones. The first zone is called forming zone S_f, which is the zone where CAVs and HDVs transition from a disordered spatial distribution to a designated spatial distribution. Afterward, the mixed vehicle platoon is formed. The numbers of CAVs and HDVs in the forming zone are N_c and N_h, respectively, and total number of vehicles is $N = N_c + N_h$. Additionally, the vehicles in each lane are indexed as $j = 1, 2, \cdots$, according to the distance from near to far the end of the forming zone. Then, all vehicles are labeled with a unique number, with the lane and position serial number, and the spatial distribution of the vehicles in each lane can be determined for further analysis of platoon formation. The second zone is cruising zone S_c, where the mixed vehicle platoon is cruising after the designated platoon formation is formed in the forming zone.

The lane-changing operation of CAVs is directly tied to the safety and efficiency of mixed vehicle platoon forming control [29], as it is influenced by surrounding vehicles and road geometry. In multi-lane traffic scenarios, the choice of the target roadway for platoon forming impacts the overall performance of mixed traffic flow. We have chosen the middle lanes for platoon forming to avoid continuous lane changes by CAVs, thus enhancing the safety and efficiency of the process. However, the target roadway can be adjusted to other lanes depending on specific traffic conditions. Additionally, we assume that HDVs are willing to accommodate CAVs merging into the same lane to form a platoon. This assumption is reasonable in highway traffic, where lane changes are relatively infrequent [30]. By doing so, it mitigates disruptive lane-changing behaviors of HDVs that could interfere with the platoon forming process.

B. Vehicle Kinematics

It is necessary for CAVs to carry out lane-keeping or lane-changing operations in platoon forming control. The bicycle model can be employed to characterize the longitudinal and lateral motion of vehicles, as shown in Eqs. (1) and (2). This model simplifies vehicle dynamics by reducing the complexity of a four-wheeled vehicle to a two-wheeled representation, consisting of a front wheel and a rear wheel. It is widely used in vehicle dynamics research due to its general applicability and reduced complexity [31, 32].

$$\dot{x}(t) = v(t)\cos\big(\theta(t) + \beta(t)\big) \tag{1}$$

$$\dot{y}(t) = v(t)\sin\big(\theta(t) + \beta(t)\big) \tag{2}$$

with

$$\dot{\theta}(t) = \frac{v(t)\sin\beta(t)}{L_r}$$

$$\beta(t) = \arctan\frac{L_r\tan\delta(t)}{L_f + L_r}$$

where v is the speed, θ is the heading angle, β is the sideslip angle at the vehicle mass center, and δ is the steering angle of the front wheel. x and y are the longitudinal and lateral

positions, respectively. L_f and L_r are the distances from the vehicle mass center to the front and rear axles, respectively.

C. Fundamental of RL

RL constitutes a pivotal branch of machine learning, which has been widely used in the field of vehicle intelligent decision-making [4, 11]. The framework of RL is formulated using the Markov Decision Process (MDP) [33], which enables an agent to learn behavior through trial-and-error interactions with a dynamic environment. During each time t of MDP, the agent observes the state $s(t)$ and chooses an action $a(t)$ based on its policy $\pi(a(t)|s(t))$, leading to subsequent states $s'(t)$ and rewards $r(t)$. The objective of RL is to learn an optimal policy $\pi^*(s(t))$ that maximizes the expected cumulative reward from each $s(t)$ over all possible policies [34]. The value of $s(t)$ under the $\pi^*(s(t))$ is given by the optimal state value function $V^*(s(t))$, and the value of taking an action under the π^* is given by the optimal action-value function $Q^*(s(t), a(t))$. The $V^*(s(t))$ and $Q^*(s(t), a(t))$ are calculated by Eqs. (3) and (4).

$$V^*\big(s(t)\big) = \max_a Q^*(s(t), a(t)) \tag{3}$$

$$Q^*\big(s(t), a(t)\big) = \tag{4}$$

$$R(s(t), a(t)) + \gamma \sum_{s'(t)} P(s'(t)|s(t), a(t)) V^*(s'(t))$$

where $\gamma \in (0,1]$ is the discount factor, $R(s(t), a(t))$ is the immediate reward received after taking action $a(t)$ in state $s(t)$, $P(s'(t)|s(t), a(t))$ is the transition probability that $s(t)$ transformed to $s'(t)$ after taking action $a(t)$. Given the optimal action-value function $Q^*\big(s(t), a(t)\big)$, the $\pi^*(s(t))$ can be derived by selecting the action that maximizes $Q^*\big(s(t), a(t)\big)$, i.e.,

$$\pi^*\big(s(t)\big) = \operatorname*{argmax}_a Q^*\big(s(t), a(t)\big) \tag{5}$$

Each CAV acts as an agent in the mixed vehicle platoon forming control. The state variables include the position and speed of the CAV and surrounding vehicles. The action variables are multiple vehicle operations, i.e., acceleration, deceleration, left lane-changing, right lane-changing, and cruising. A large number of state variables and actions makes it extremely sophisticated for model-based RL methods to solve this multi-agent control problem.

In addressing the difficulty in the multi-agent and decentralized platoon forming control, the model-free Advantage Actor-Critic (A2C) algorithm is used in the RL. The A2C algorithm applies to decentralized systems since it employs both an individual actor network (policy) and a critic network (value) to optimize the learning process, where the actor determines the actions to be taken given the current state, and the critic evaluates the performance of the chosen actions. A2C utilizes neural networks $\pi_\sigma\big(a(t)\big|s(t)\big)$ for policy representation, where σ is the learnable parameters [35], which is updated by optimizing the objective function Eq. (6).

$$J(\theta) = \mathbb{E}[\log \pi_\sigma(a(t)|s(t)) \cdot \mathcal{A}(s(t), a(t))] \tag{6}$$

(a) Uniform distribution.

(b) Centralized distribution.

(c) Random distribution.

Fig. 4. Three typical formations of a mixed vehicle platoon.

where $\mathcal{A}(s(t), a(t)) = Q^{\pi_\sigma}\big(s(t), a(t)\big) - V_\omega^{\pi_\sigma}\big(s(t)\big)$ is the advantage function to update the weights, which can significantly enhance the stability of the training process [36].

A2C has several limitations, including potential biased gradient estimates due to bootstrapping in the critic's value function, sensitivity to hyperparameters leading to instability, and scalability issues with synchronous updates. High variance in gradient estimates can also affect learning stability. These problems can affect the training efficiency and success rate of the CAV control policy for platoon forming. To address these issues, this study employs experience replay to reduce biased gradients [37], uses grid search to optimize hyperparameters [38], adopts asynchronous variants to improve scalability [39], and applies entropy regularization to encourage exploration and stabilize learning [40].

IV. Mixed Vehicle Platoon Forming Method

A. Mixed Vehicle Platoon Formation Generation

CAVs are distinguished by over-the-horizon perception, sensitive responsiveness and almost identical driving behaviors, whereas HDVs are characterized by longer reaction times, perception errors, and a variety of driving styles. The spatial distribution of CAVs and HDVs affects the mixed vehicle platoon's performance in terms of driving safety, energy consumption, and travel time due to these differences.

Generating a feasible platoon formation is the prerequisite of platoon forming control. There are three common mixed vehicle platoon formations, i.e., uniform, centralized, and random distribution of CAVs, as shown in Fig. 4. In particular, platoon formation with uniform distribution has shown the best performance in stabilizing the traffic flow and enhancing traffic efficiency [20, 21]. This is because CAVs can guide and stabilize the driving behavior of following HDVs. In addition, according to Yao *et al*. [41], a mixed vehicle platoon in which the lead vehicle is a CAV will improve the platoon's energy and traffic efficiency because of the controllability of CAVs.

The traffic states (i.e., the original spatial distribution and movement of CAVs and HDVs) must also be considered while generating platoon formation. Four priority-based guidelines (in the order of priority) are designed to determine the feasible

mixed vehicle platoon formation $\mathcal{F}$ based on the empirical formation method. This method accounts for vehicle states to ensure collision avoidance while adhering to the guidelines of platoon operation. It leverages the capabilities of CAVs to guide HDVs toward energy-efficient driving [42].

1) A CAV prefers merging to the nearest possible position in the platoon formation [43];

2) A CAV prefers staying ahead of at least one HDV in the platoon formation [44];

3) The leader vehicle in the mixed platoon is a CAV [44];

4) The CAVs are distributed uniformly within the mixed vehicle platoon [21].

Based on the above guidelines, we propose the mixed vehicle platoon be formed through the following three steps:

Step 1: Searching the feasible spaces and sequence numbers for forming the mixed vehicle platoon. The number of CAVs, the number of vehicles in the target lane where the mixed vehicle platoon is formed, and the intervehicle headway in the target lane are used to derive the feasible spaces and numbers for forming the platoon. We assume the speed of HDVs is constant in every time step during the search process, given that the time step is small and vehicle speed fluctuations are minimal over short durations in highway traffic [11, 45]. More specifically, the feasible space is behind the last vehicle and in front of the lead vehicle. Within the vehicle group, a space is feasible when the headway exceeds the maximum headway T_{hmax}. According to guidelines 1)-4), the feasible space is selected only from the front to the back based on the number of CAVs. Then, feasible space numbers are determined by using the sequence number of the mixed platoon.

Step 2: Calculating the priority of each CAV merges into each feasible space. The priority of each CAV determines which one merges into the specific space. As guideline 1), the priority assigned to every feasible space is ascertained by measuring the distance of each CAV from the feasible space; the greater the distance, the lower the priority. Furthermore, for guidelines 2) and 3) that are met, the priority is raised by one to attain uniform distribution and leader CAV formation. Finally, the priority sequence corresponding to all CAVs for each feasible space can be obtained.

Step 3: Ensuring the feasible mixed vehicle platoon formation. The CAVs with the highest priority are arranged at feasible spaces in order from front to back in the target lane. Combined with the original vehicle positions in the target lane and newly arranged CAVs, then $\mathcal{F}$ can be determined.

The implementation of the mixed vehicle platoon formation generation based on the above three steps is summarized in Algorithm 1. In this algorithm, x_z v_z, $T_{h,z}$ are the longitudinal position, longitudinal speed, and headway of zth vehicle from front to back in the target lane, respectively. N_l is the current number of vehicles in the target lane, $\mathcal{H}$ is the feasible space sequence set which stores the sequence number of spaces in the platoon, $Num(\mathcal{H})$ is the number of elements in $\mathcal{H}$, $\hbar_m$ is the mth element from front to back in $\mathcal{H}$, x_m is the longitudinal position of mth space, d_{km} is the longitudinal distance between kth CAV and mth space, $\wp_{km}$ is the priority of the kth CAV for the mth space, λ_{km} is the indices obtained by sorting d_{km} of all CAVs at mth space in descending order, μ_{km} is the indices obtained by sorting x_k of all CAVs in ascending order, n_m is the number of HDVs following mth space in target lane.

After determining the $\mathcal{F}$, each CAV's safe and effective driving control for platoon forming is thoroughly explained in Section IV. B. Note that the $\mathcal{F}$ is calculated every five simulation steps until all CAVs merge into the target lane to form a mixed vehicle platoon, to account for the impacts of random driving behaviors of HDVs.

Algorithm 1 Mixed vehicle platoon formation generation

Input: N_c, N_l, x_z v_z, and $T_{h,z}$

Output: $\mathcal{F}$

Initialize $\mathcal{H} \leftarrow [1]$
for $z \leftarrow 2$ to N_l **do**
  Calculate $T_{h,z}$
  If $T_{h,z} \geq T_{hmax}$ and $Num(\mathcal{H}) < N_c$ **do**
    $\mathcal{H} \leftarrow \mathcal{H} \cup [z]$
  end if
end for
If $Num(\mathcal{H}) = N_c - 1$ **do**
  $\mathcal{H} \leftarrow \mathcal{H} \cup [N_l + 1]$
end if
for each $\hbar_m$ in $\mathcal{H}$ **do**
  Extract x_m for mth space labeled by $\mathcal{H}$
  for $k \leftarrow 1$ to N_c **do**
    $\wp_{km} \leftarrow 0$
    Calculate d_{km}
  end for
  Sort d_{km} and update $\wp_{km} \leftarrow \wp_{km} + \lambda_{km}$
  if mth space is followed by HDV **do**
    Update $\wp_{km} \leftarrow \wp_{km} + \lambda_{km} + n_m$
  end if
  if mth space is the leader **do**
    Sort x_k and update $\wp_{km} \leftarrow \wp_{km} + \mu_{km}$
  end if
  Select the CAV with the highest $\wp_{km}$ as the feasible occupant of the mth space in the platoon, and update $\mathcal{F}$
end for
return $\mathcal{F}$

B. Mixed Vehicle Platoon Forming Control

The CAV's lane-changing trajectory and speed profile are controlled for it to move from its original position to the target position at the right moment, and then form the designated platoon formation with the HDVs. Given that multiple CAVs often need to be controlled and highly random and unpredictable driving behaviors of HDVs in mixed vehicle platoon forming, a MARL framework is developed for mixed vehicle forming control with multiple CAVs. The MARL

framework is defined as a partially observable MDP, where each agent (i.e., CAV) can only observe part of the states from the surrounding traffic environment to reduce the redundancy of states and improve learning efficiency. This is because the forming control for CAVs is closely related to surrounding vehicle states. Then, each agent follows a decentralized policy to choose a at time t, achieve safe and efficient forming. The definition of the partially observable MDP is as follows:

1) State Space: The state space of agent k is $\mathcal{s}_k$, which is a 6×4 matrix. The subscript $k = 1, 2, \cdots, N_c$ indicates the vehicle number of CAVs. The number of rows in the $\mathcal{s}_k$ is defined by considering the agent k and the five nearest vehicles around it. The number of columns is defined by considering the position and speed both longitudinally and laterally of each vehicle. Then, the first row of $\mathcal{s}_k$ represents the states of the agent k and the remaining five rows represent the states of the five nearest vehicles.

2) Action Space: The action space $\mathcal{a}_k$ of agent k includes five operations based on the vehicle dynamics: accelerating, braking, cruising, changing lanes to the left, and changing lanes to the right. The maximum acceleration and deceleration are used for accelerating and braking at each step respectively, to simplify system complexity and facilitate efficient training in a discrete-time environment. With a time step of 0.1s in the simulation, using fixed maximum values for acceleration and deceleration is reasonable based on the designs in [46]. This small interval allows the model to make frequent adjustments, enabling nuanced control over longer periods. When cruising or changing lanes, the vehicle speed remains constant from the start of the operation. The action space combination for all CAVs is defined as $a = a_1 \times a_2 \times \cdots \times a_{N_c}$.

3) Reward Function: The reward function covers multiple objectives including vehicle driving safety, traffic efficiency, energy efficiency, and success of platoon formed. Then, the reward $r_k(t)$ for the agent k at time t is defined in Eq. (7).

$$r_k(t) = \varpi_c r_{c,k}(t) + \varpi_h r_{h,k}(t) + \varpi_s r_{s,k}(t) + \varpi_e r_{e,k}(t) + \varpi_p r_{p,k}(t) \tag{7}$$

where ϖ_c, ϖ_h, ϖ_s, ϖ_e, and ϖ_p are weighting coefficients for avoiding collision, keeping safe headway, ensuring suitable speed, reducing energy consumption, and promoting platoon forming, respectively. Each reward is defined as follows.

Collision avoidance reward $r_{c,k}(t)$: CAVs need to avoid collisions with surrounding vehicles while traveling for platoon formatting, and the $r_{c,k}(t)$ is defined in Eq. (8).

$$r_{c,k}(t) = \begin{cases} -1 & \text{collision} \\ 0 & \text{safe} \end{cases} \tag{8}$$

Headway keeping reward $r_{h,k}(t)$: A small time headway increases the risk of collision between the CAV and the preceding vehicle, while a large time headway reduces traffic capacity. Then, the $r_h(t)$ is defined in Eq. (9).

$$r_{h,k}(t) = \log \frac{T_{h,k}(t)}{T_{hmin}} \tag{9}$$

where $T_{h,k}(t)$ is the headway between the kth CAV and its preceding vehicle. T_{hmin} is the minimum time headway.

Suitable speed ensuring reward $r_{s,k}(t)$: The traffic efficiency is determined by the speed at which a CAV travels; if CAV moves too slowly, it will not only lengthen its travel time but will also cause delays for the following vehicles. To stimulate the CAV travel at high speed, the $r_{s,k}(t)$ is

$$r_{s,k}(t) = \min\left\{\frac{v_k(t) - V_{min}}{V_{max} - V_{min}}, 1\right\} \tag{10}$$

Energy consumption reduction reward $r_{e,k}(t)$: Less energy consumption is expected to improve vehicle energy efficiency, leading to form the sustainable transportation, where energy efficiency is defined by comparing energy consumption of the CAV to energy consumption of average traffic flow. Then, the $r_{e,k}(t)$ is defined in Eq. (11).

$$r_{e,k} = \frac{E_{ref}(t) - E_k(t)}{E_{ref}(t)} \tag{11}$$

where E_{ref} is the reference energy consumption, which is calculated using the average traffic flow speed.

Platoon forming promoting reward $r_{p,k}(t)$: This reward is used to promote platoon forming, which depends on where the CAV is located inside the designated formation. Note that if kth CAV not in the target lane, $r_{p,k}(t) = 0$. If the kth CAV is both the designated and actual leader vehicle, then $r_{p,k}(t)$ is

$$r_{p,k}(t) = \begin{cases} 2 & T_{hmin} \le T_{h,k}(t) < T_{hmax} \\ 1 & \text{others} \end{cases} \tag{12}$$

If the kth CAV is the follower vehicle and in its designated position in the platoon, the reward is defined with the consideration of the platoon sequence and CAV position, i.e.,

$$r_{p,k}(t) = r_{ps,k}(t) + r_{pp,k}(t) \tag{13}$$

where $r_{ps,k}(t)$ and $r_{pp,k}(t)$ are the rewards benefit from the platoon sequence and the CAV position, respectively. The $r_{ps,k}(t)$ and $r_{pp,k}(t)$ are given by

$$r_{ps,k}(t) = \begin{cases} 1 & \text{if } \vartheta_{p,k}(t) = 0 \text{ and } \vartheta_{r,k}(t) = 0 \\ 0 & \text{if } \vartheta_{p,k}(t) = 1 \text{ and } \vartheta_{r,k}(t) = 0 \\ -1 & \text{others} \end{cases} \tag{14}$$

$$r_{pp,k}(t) = \begin{cases} 1 & T_{hmin} \le T_{h,k}(t) < T_{hmax} \\ 0 & \text{others} \end{cases} \tag{15}$$

where $\vartheta_{p,k}(t)$ and $\vartheta_{r,k}(t)$ are the preceding and rear vehicle type flags of the kth CAV, respectively. A value of 1 indicates that the vehicle is a CAV, while a value of 0 indicates the HDV.

It should be noted that vehicle driving safety is achieved using collision avoidance and safe headway-keeping rewards. The traffic efficiency is ensured using suitable speed ensuring reward, while the vehicle energy efficiency is improved using energy consumption reduction reward. The designated formation is successfully formed using platoon forming promoting reward.

4) Safety Supervisor: The process of RL training involves some unavoidably risky actions of CAVs, which impact the speed and success of RL training. In this context, we define three safe driving rules: 1) Prohibiting two or more consecutive lane-changing operations; 2) Considering the states of surrounding vehicles in action decision to avoid rear-end and

side-side collisions; 3) Avoiding heavy braking or acceleration when the CAV is too close to the preceding and rear vehicles. Filtering the calculated actions a_k using safe driving rules and historical action $a_{h,k}$ yields supervised action $a_{s,k}$. For clarity, the implementation algorithm for the safety supervisor is given in Algorithm 2.

Algorithm 2 Safety supervisor

Input: N_c, $T_{h,k}$, a_k, and $a_{h,k}$

Output: $a_{s,k}$

Initialize $k \leftarrow 1$, $a_{s,k} \leftarrow a_k$
while $k \leq N_c$ **do**
  if a_k is lane-changing and $a_k \in a_{h,k}$ **then**
    $a_{s,k} \leftarrow$ cruising
  else
    if $T_{h,k} < T_{hmin}$ **then**
    $a_{s,k} \leftarrow$ braking
    end if
  end if
  $k \leftarrow k + 1$
end while
return $a_{s,k}$

Algorithm 3 Mixed vehicle platoon forming

Input: $\mathcal{M}$, $\mathcal{T}$, $\mathcal{D}$, and $\mathcal{F}$

Output: $\pi_\sigma(s(t))$

for episode $\leftarrow$ 1 to $\mathcal{M}$ **do**
  Initialize $\mathcal{D} \leftarrow \{\}$, $t \leftarrow 0$
  while $t \leq \mathcal{T}$ **do**
  Calculate the $\mathcal{F}$ using Algorithm 1
    for $k \leftarrow 1$ to N_c **do**
      Calculate $a_k(t)$ using $\pi_\sigma(s(t))$
      Update $a_{s,k}(t)$ using Algorithm 2
      kth Agent executes $a_{s,k}(t)$
      Calculate $r_k(t)$ based on $\mathcal{F}$, $a_{s,k}(t)$ and $s(t)$
      Transition to $s_k(t+1)$
    end for
  $t \leftarrow t + \Delta t$
  end while
  Calculate $Q^{\pi_\sigma}(s(t), a(t))$ and $\mathcal{A}(s(t), a(t))$
  Store $\{s(t), a_{s,k}(t), \mathcal{F}, Q^{\pi_\sigma}(s(t), a(t)), \mathcal{A}(s(t), a(t))\}$ into $\mathcal{D}$
  Shuffle the order of $\mathcal{D}$
  Select different groups of $\mathcal{D}$ to update the policy and state value networks, resulting in updated $\pi_\sigma(s(t))$ for each agent
end for
return $\pi_\sigma(s(t))$

5) Network Structure: The A2C network of each agent is shown in Fig. 5. The inputs of the A2C network include S_i and the designated mixed vehicle platoon formation, while the output is actions of all agents. In the A2C network, each type of input divided by the physical definition (i.e., vehicle longitudinal and lateral speed and position) is firstly fed into one 64-neuron fully connected (FC) network. The 5 independent FC networks form the hidden layer. Subsequently, all outputs in the hidden layer are combined and fed into the 128-neuron FC network. Then the actor-critic network will update the policy and value with the learned features.

We are setting each agent to have the same A2C network structure, to formulate the multi-agent network. As shown in Fig. 6, each agent has its actor network that makes decisions based on a combination of local observations and actions from all the agents. The critic network is centralized and has access to the actions and states of all agents, enabling it to evaluate the joint action-value function. This approach, known as Centralized Training with Decentralized Execution [47], can accommodate both the complex training process of MARL and the autonomy in execution by individual agents. The mixed vehicle platoon forming method is summarized in Algorithm 3, where $\mathcal{M}$ is the number of episodes, $\mathcal{T}$ is the number of policy steps per episode, and $\mathcal{D}$ is the memory buffer. Each agent selects an action using the shared policy, updates actions according to the safety supervisor, and accumulates experience such as the states, chosen actions, and feasible platoon formation at each policy step to update the policy.

V. Simulation and Results

To verify the performance and necessity of forming mixed vehicle platoons, several simulations are conducted using MATLAB (version 9.14, 2023a) on a computer with an Intel Core i7-13700KF @ 3.40 GHz CPU, 64GB RAM, and NVIDIA RTX 4080 GPU. First, the parameters for traffic scenarios and typical vehicles are defined. Next, various MARL algorithms are compared to find the best one for developing the mixed vehicle platoon forming control policy. Finally, the effectiveness of the proposed method is evaluated based on the platoon formation success rate and improvements in energy and traffic efficiencies.

A. Traffic Scenario and Vehicle Setup

1) Traffic Scenario: The highway traffic is constructed using the Highway-env simulator [27], to test the effectiveness of the proposed method. This simulator offers extensive flexibility in configuring highway scenarios to meet experimental needs. It provides access to a variety of traffic information and supports integrated vehicle and traffic decision-making processes, making it widely applied in studies of CAVs and mixed traffic control using RL [11].

The parameters of the traffic scenario are listed in Table I. In the forming zone, three CAVs with randomly dispersed initial positions and lanes within 100m of the starting position. There are also three HDVs distributed in the center lane randomly. Other surrounding HDVs are randomly distributed on the road. In addition, the initial speed of each vehicle is randomly distributed around the average traffic flow speed. Note that the chosen number of vehicles and CAVs is representative. A 50% CAV penetration rate maximizes mixed vehicle platoon configurations, thoroughly validating the proposed method's performance. The number of vehicles in the platoon, which

directly affects computational complexity, is determined based on the capabilities of our simulation computer. Naturally, the proposed method remains applicable to both larger and smaller platoon sizes.

2) Typical Vehicles: The vehicle's energy consumption is calculated using the widely accepted wheel-to-distance model [48], as expressed in Eq. (16).

$$E_e = (mgf\cos\psi + 0.5C_dA\rho v^2 + mg\sin\psi + m\dot{v})v\Delta t \quad (16)$$

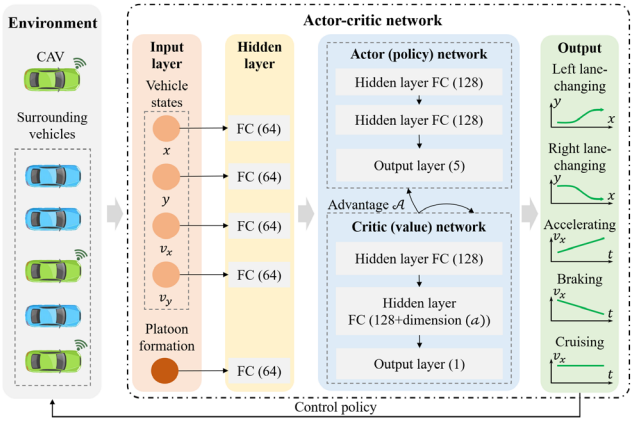


Fig. 5. Actor-critic network and its interaction with environment.

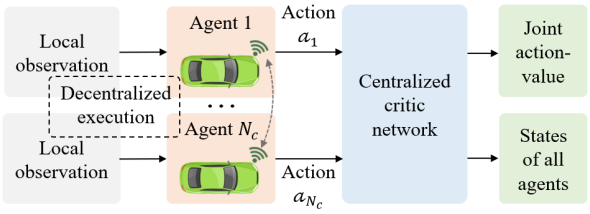


Fig. 6. Multi-agent network.

TABLE I
TRAFFIC SCENARIO PARAMETERS

Parameter	Value	Parameter	Value
Average traffic flow speed	25m/s	Minimum speed V_{min}	5m/s
Cruising zone length S_c	9400m	Number of CAVs N_c	3
Forming zone length S_f	600m	Number of HDVs N_h	8-10
Maximum speed V_{max}	33m/s	Width of lane D_l	4m

TABLE II
TYPICAL VEHICLE PARAMETERS

Parameters	Type 1	Type 2	Type 3	Type 4	Type 5	Type 6
Vehicle mass m	1545kg	1015kg	1375kg	1430kg	1067kg	1155kg
Body length	5.21m	3.85m	4.23m	4.25m	3.92m	4.03m
Body width	2.04m	1.71m	1.98m	2.10m	1.78m	1.83m
Rolling resistance f	0.020	0.022	0.019	0.021	0.023	0.024
Frontal area A	2.33m^2	2.19m^2	2.40m^2	2.46m^2	2.14m^2	2.04m^2
Air drag coefficient C_d	0.31	0.33	0.29	0.37	0.33	0.32

TABLE III
TRAFFIC SCENARIO PARAMETERS

Parameter	Value	Parameter	Value
Batch size	64	Learning rate	0.0002
Collision avoiding weighting ϖ_c	200	Platoon forming weighting ϖ_p	2
Energy reduction weighting ϖ_e	1	Suitable speed ensuring weighting ϖ_s	1
Headway keeping weighting ϖ_h	1	Time discount factor	0.99

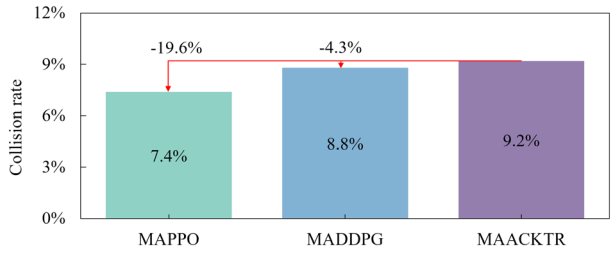


Fig. 7. Collision rate of mixed vehicle platoon forming.

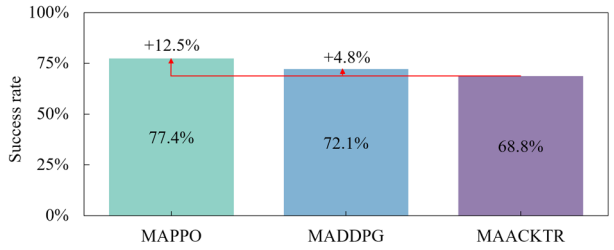


Fig. 8. Success rate of mixed vehicle platoon forming.

where m is the vehicle mass, f is the rolling resistance coefficient, C_D is the aerodynamic drag coefficient, A is the frontal area, ρ is the air density, and ψ is the road slope.

Given the variety of vehicle types in real-world traffic, we selected six typical vehicles to simulate the platoon heterogeneous, encompassing the most common passenger vehicle categories found in real-world traffic [49]. The parameters of six typical vehicles are listed in Table II. In addition, the maximum deceleration and acceleration are -4 m/s^2 and 4 m/s^2, respectively. The range of steering angle is from -1rad to 1rad. The minimum and maximum headway are 0.8s and 2s, respectively. Note that the type of each vehicle is determined by a random distribution of six typical vehicles. The car-following behavior of HDVs is simulated using the Intelligent Driver Model [50] with random sampling in the distributions of headway. The lane-changing behavior is captured by using the Microscopic Online Behavior Model [51].

B. MARL Setup and Evaluation

1) MARL Setup: The traffic scenario definition states that three CAVs must merge into the target lane and form the mixed vehicle platoon with HDVs. Consequently, there are three agents in MARL. The parameters of MARL are listed in Table III. Note that if an episode is completed or a collision occurs, the traffic environment will be reset to its initial state and change random seeds to start a new epoch.

2) Evaluation of Different MARL Algorithms: To identify a suitable MARL algorithm for mixed vehicle platoon forming control, three candidates are evaluated: Multi-agent Proximal Policy Optimization (MAPPO) [52], Multi-agent Actor-critic using Kronecker-Factored Trust Region (MAACKTR) [53], and Multi-agent Deep Deterministic Policy Gradient (MADDPG) [54]. Three candidates are evaluated with the same traffic environment, reward, and hyper-parameter settings. The collision rate and success rate of the platoon forming control after about 10000 episodes are shown in Figs. 7 and 8, respectively. Note that the success rate in this context includes collision incidents, where any collision is considered a failure.

In Fig. 7, using the MAACKTR during RL training results in the highest collision rate, i.e., up to 9.2%. Compared to MAACKTR, the collision rates of MAPPO and MADDPG decreased by 19.6% and 4.3%, respectively. This suggests that the MAPPO yields the lowest collision rate with advantages in efficient exploration under safe driving concerns in dynamic traffic. Note that the collision rate is not zero because these tests did not include the safety supervisor model. This model adds an extra layer of safety by monitoring and intervening to prevent collisions. Without it, the algorithms rely solely on their learned policies to avoid collisions, resulting in higher collision rates. In addition, using the MADDPG results in a 68.8% platoon-forming success rate, as illustrated in Fig. 8. The highest success rate is 77.4% when the MAPPO is employed. In summary, MAPPO improves mixed vehicle platoon forming success rate while guaranteeing driving safety, which performs better than MAACKTR and MADDPG. Therefore, MAPPO is used to train the forming control policy for CAVs.

C. Evaluation of Platoon Forming

1) Rule-based Method: The rule-based method for mixed vehicle platoon forming is defined as a benchmark to evaluate the platoon forming success rate of the proposed method, which is designed by Maiti et al. [26]. In the rule-based method, the leading vehicle toward the movement direction is selected as the platoon leader. Other CAVs are incorporated into the platoon as followers, according to the descending order of their positions. When the requirements for driving safety are satisfied, all CAVs are encouraged to merge into the platoon.

2) Results: For the proposed method, the training consists of 40,000 episodes using MAPPO. Fig. 9 illustrates the average reward and the upper and lower limits of reward in ten repeated RL training. In Fig. 9, the reward fluctuates and changes, increases gradually, and converges to 180 finally. This shows that the RL training process is effective and reasonable.

In the comparison tests, 1000 random traffic scenarios are used for the comparative analysis, and the rule-based and proposed methods are tested separately in each traffic scenario. The CAV position and speed trajectories at one of the comparison tests are depicted in Figs 10 and 11, respectively, and the formed platoon is shown in Fig. 12. In addition, Figs. 13 (a) and 13 (b) give the distributions of platoon forming time and energy consumption, respectively. Table IV lists the success rate of platoon forming the random formation and optimal formation (i.e., uniform distribution of CAVs).

TABLE IV
RESULTS OF SUCCESS RATE

Method	Random formation	Optimal formation
Rule-based	87%	4%
Proposed	85%	11%

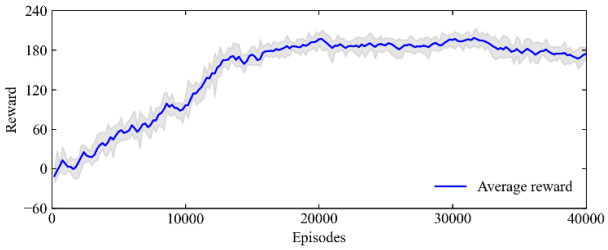


Fig. 9. The rewards and boundaries of 40,000 episodes in RL training.

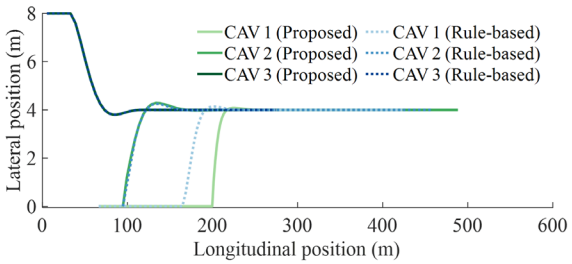


Fig. 10. The position trajectory of HDVs and CAVs.

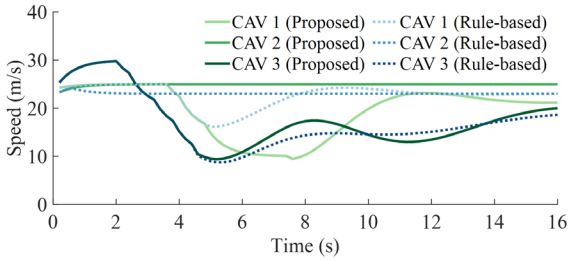


Fig. 11. The speed trajectory of HDVs and CAVs.

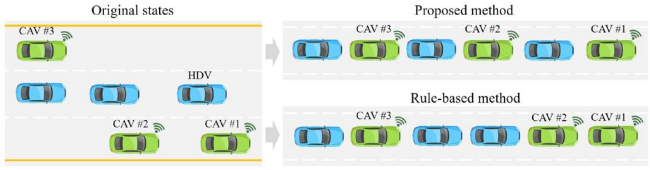
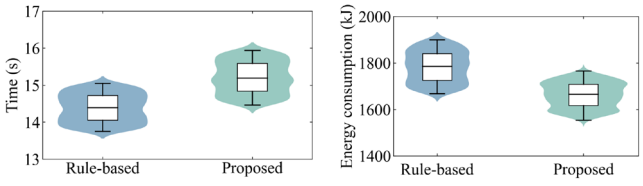


Fig. 12. The formed mixed vehicle platoon using rule-based and proposed methods.



(a) forming time (b) energy consumption

Fig. 13. The time and energy consumption of mixed vehicle platoon forming.

This article has been accepted for publication in IEEE Internet of Things Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2025.3535732

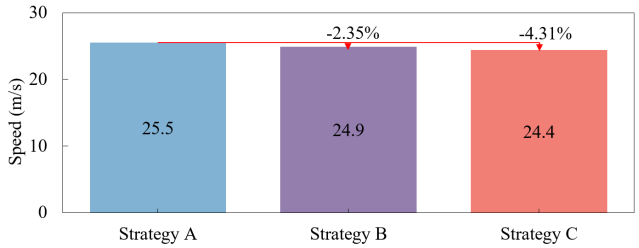


Fig. 14. The average speed of all vehicles in the forming zone.

Fig. 15. The average energy of all vehicles in the forming zone.

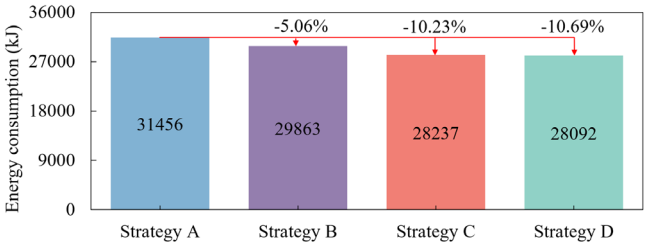


Fig. 16. The average energy of mixed vehicle platoon in whole travel.

Fig. 17. The average travel time of mixed vehicle platoon in whole travel.

As shown in Figs. 10 and 11, the CAVs started from either the leftmost or the rightmost lanes. Under the control of rule-based and proposed methods, CAVs can change lanes to merge into the middle lane and collaborate with the HDVs to form a mixed vehicle platoon. In contrast to the rule-based method, as shown in Fig. 12. which can only form a mixed vehicle platoon with the random distribution of CAVs, the proposed method can form a mixed vehicle platoon with a uniform distribution of CAVs.

The range of platoon forming time for the rule-based method is 13.6s-15.1s, and 14.2s-16.1s for the proposed method, as indicated in Fig. 13 (a). The average platoon forming time is comparable between the rule-based method (14.4s) and the proposed method (15.2s). However, the proposed method saves 7.19% in energy consumption during platoon forming compared to the rule-based method, as shown in Fig. 13 (b). This is because the speed trajectory of the proposed method is smoother than the rule-based method, as shown in Fig. 11, which improves the energy efficiency of the CAVs. Table IV shows that the proposed method increases the optimal formation achievement rate by 175% over the rule-based method, without compromising platoon forming success rate.

The above findings indicate that both the rule-based and proposed methods can rapidly control the CAVs to form the mixed vehicle platoon in dynamic traffic. However, the proposed method is more energy-efficient, while having a higher success rate in forming optimal formations.

D. Evaluation of Energy and Traffic Efficiencies

The investigation involves 1000 individual simulation trials with random traffic scenarios to analyze the impact of the proposed method on energy and traffic efficiency. This evaluation covers both the platoon forming zone and the whole travel including forming and cruising zones.

1) Cruising Speed Optimization: Once the mixed vehicle platoon has formed the designated formation, the driving speed of the mixed vehicle platoon has an impact on the platoon's total energy consumption and travel time. The leader CAV speed optimization and the car-following speed control of followers are the two parts of cruising speed control. The trigonometric speed profile is used to derive the vehicle energy-saving driving speed of the leader CAV in a mixed vehicle platoon, which is widely used in vehicle eco-driving control [31, 32].

The car-following behaviors between vehicles in a platoon can be categorized into three types based on the kind of leading and trailing vehicles, i.e., HDV-HDV, HDV-CAV, CAV-HDV, and CAV-CAV. If both the preceding and following vehicles are HDVs or the preceding vehicle is a CAV and the following vehicle is an HDV, the HDVs' car-following behavior is captured using the intelligent driver model [50]. The adaptive cruising control strategy [41] is employed to identify CAV's car-following behavior in the CAV-HDV type. If both the preceding and following vehicles are CAVs, the cooperative adaptive cruising control strategy [41] is utilized to simulate the car-following behavior of the following CAV.

2) Strategy Design: To evaluate the energy and traffic efficiencies of the proposed method, four strategies, named Strategies A, B, C, and D, are compared, where the proposed forming method is used in Strategies C and D. Platoon forming control is not considered in Strategies A and B. All CAVs and HDVs driving freely in Strategy A. In Strategy B, the CAVs use the same trigonometric speed profile model [31] to derive energy-saving speed as the proposed method, but the HDVs are free to drive. In Strategies C and D, the platoon forming control is identical to the proposed method, but the leader vehicle in the mixed vehicle platoon drives at the constant average traffic flow speed in Strategy C and drives at optimized speed using the trigonometric speed profile model in Strategy D. Furthermore, when the preceding vehicle is encountered in Strategies A, B, C, and D, the car-following behavior is controlled via the intelligent driver model [50], adaptive cruising, and cooperative adaptive cruising control strategies [41].

3) Results of Forming Zone: To evaluate the impact of CAVs on the energy and traffic efficiencies of surrounding vehicles

during the mixed platoon forming, the average speed and energy consumption of all vehicles in the forming zone are collected, as shown in Figs. 14 and 15, respectively. Note that Strategy D is not included in this evaluation since Strategies C and D share the same platoon forming strategy. The percentages of Strategies B and C are shown in Figs. 14 and 15 are calculated as the relative reductions in average speed and energy consumption compared to Strategy A.

As shown in Fig. 14, compared to Strategy A, the average speed of all vehicles decreases by 2.35% under Strategy B due to the optimized speed profile, and by 4.31% under Strategy C due to the control of mixed vehicle platoon formation. In Fig. 15, Strategy C improves energy efficiency by an average of 1.3% and Strategy B improves by an average of 4.71% compared to Strategy A. The results indicate that during the forming of mixed vehicle platoon, CAVs need to change lanes under control, which affects the movement of surrounding vehicles. However, the impact on surrounding vehicles is not significant, so the average speed and energy differences among all vehicles are not substantial.

4) Results of the Whole Travel: The average energy consumption and travel time for 1000 individual simulations of a mixed vehicle platoon throughout the whole travel, are displayed in Figs. 16 and 17, respectively. Note that the energy consumption here refers to the sum of the energy of the six vehicles planned to form the mixed vehicle platoon, and the travel time is the time taken by the last of these vehicles to arrive at the destination. The percentages of Strategies B, C, and D in Figs. 16 and 17 are calculated as the relative reductions in average energy consumption and time compared to Strategy A.

As shown in Fig. 16, the energy consumption of Strategy B is reduced by 5.06% compared to Strategy A. This indicates that the CAV drives at optimized vehicle speeds can improve the energy efficiency of mixed traffic. However, when forming the mixed vehicle platoon, Strategy C improves the holistic energy efficiency of the mixed vehicle platoon by 10.23% and 5.44% compared to Strategies A and B, respectively. This indicates that forming a mixed vehicle platoon can effectively improve the vehicle energy efficiency. In addition, in Strategy D, thanks to the adoption of the optimized speed of leader CAV, the energy consumption is further reduced by 0.51% compared to Strategy C. Nonetheless, there is a minor improvement in travel time reduction between Strategies C and D., as displayed in Fig. 17, forming a mixed vehicle platoon contributes to the improvement of traffic efficiency, and the time-saving performance of the mixed vehicle platoon is further improved when the optimized speed control is employed.

In summary, it is highly advantageous to enhance the energy and traffic efficiency than disorganized mixed traffic by controlling CAVs in conjunction with HDVs to create a mixed vehicle platoon. Furthermore, properly regulating the speed profile of CAVs in platoons of mixed vehicles could improve overall energy performance and traffic efficiency, which maximizes the benefit of forming a mixed vehicle platoon from random traffic flow.

VI. Conclusion

To energize the potential of CAVs in traffic safety, economy, and efficiency improvement for mixed traffic flows, this study proposes the mixed vehicle platoon forming method that considers the dynamic HDVs and CAVs driving behaviors. It uses a two-stage hierarchical control framework to realize safe and efficient platoon forming control in dynamic mixed traffic. The mixed platoon formation generation stage creates the feasible formation appropriate for the traffic scenario based on the empirical formation method and the states of CAVs and HDVs, which is the guidance of the platoon forming control. The MARL is used in the second stage for realizing the platoon forming control efficiently. By the formation requirements, CAVs in different lanes perform lane-changing operations under the control of the MARL to drive into the designated positions in the mixed vehicle platoon.

Extensive simulations are conducted using the Highway-env simulator to evaluate the effectiveness of the proposed method. The platoon forming success rate evaluation results show that the proposed method can rapidly control the CAVs to form the designated formation in dynamic mixed traffic while saving vehicle energy. Furthermore, it is extremely beneficial to control CAVs in tandem with HDVs to form a mixed vehicle platoon, which will improve energy and traffic efficiency compared to disorderly mixed traffic.

Future research will focus on integrating collision avoidance mechanisms and dynamic lane-changing behaviors of HDVs into mixed vehicle platoon forming. This integration aims to enhance the adaptability of the design method to more complex traffic scenarios, such as urban traffic and dense traffic flows.

References

[1] A. Matin and H. Dia, “Impacts of connected and automated vehicles on road safety and efficiency: a systematic literature review,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 3, pp. 2705-2736, Mar. 2023.

[2] H. Dong, W. Zhuang, B. Chen, G. Yin, and Y. Wang, “Enhanced eco-approach control of connected electric vehicles at signalized intersection with queue discharge prediction,” *IEEE Transactions on Vehicular Technology*, vol. 70, no. 6, pp. 5457-5469, June 2021.

[3] H. Dong, W. Zhuang, B. Chen, Y. Lu, S. Liu, S. Liu, L. Xun, D. Pi, and G. Yin. “Predictive energy-efficient driving strategy design of connected electric vehicle among multiple signalized intersections,” *Transportation Research Part C: Emerging Technologies*, vol.137, pp. 103595, Apr. 2022.

[4] J. Zhan, Z. Ma, and L. Zhang, “Data-driven modeling and distributed predictive control of mixed vehicle platoons," *IEEE Transactions on Intelligent Vehicles*, vol. 8, no. 1, pp. 572-582, Jan. 2023.

[5] Q. Li, Z. Chen, and X. Li, “A review of connected and automated vehicle platoon merging and splitting operations,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 12, pp. 22790-22806, Dec. 2022.

[6] Y. Wu, D. Wang, and F. Zhu, “Influence of CAVs platooning on intersection capacity under mixed traffic,” *Physica A: Statistical Mechanics and its Applications*, vol. 593, pp. 126989, May 2022.

[7] S. Gong and L. Du, “Cooperative platoon control for a mixed traffic flow including human drive vehicles and connected and autonomous vehicles,” *Transportation Research Part B: Methodological*, vol. 116, pp. 25-61, Oct. 2018.

[8] T. Miqdady, R. de Oña, J. Casas, and J. de Oña, “Studying traffic safety during the transition period between manual driving and autonomous driving: a simulation-based approach,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 6, pp. 6690-6710, June 2023.

[9] S. Mousavi, S. Bahrami, and A. Kouvelas, “Controller design for a mixed traffic system travelling at different desired speeds,” *European Journal of Control*, vol. 68, pp. 100698, Nov. 2022.

[10] Y. Xue, X. Zhang, Z. Cui, B. Yu, and K. Gao, “A platoon-based cooperative optimal control for connected autonomous vehicles at highway on-ramps under heavy traffic,” *Transportation Research Part C:*

Emerging Technologies, vol. 150, pp. 104083, May 2023.

[11] Q. Wang, H. Dong, F. Ju, W. Zhuang, C. Lv, L. Wang, and Z. Song, "Adaptive leading cruise control in mixed traffic considering human behavioral diversity," *IEEE Transactions on Intelligent Transportation Systems*, vol. 25, no. 6, pp. 5059-5070, June 2024.

[12] D. Li, K. Zhang, H. Dong, Q. Wang, Z. Li, and Z. Song, "Physics-augmented data-enabled predictive control for eco-driving of mixed traffic considering diverse human behaviors," *IEEE Transactions on Control Systems Technology*, vol. 32, no. 4, pp. 1479-1486, July 2024.

[13] Y. Li, B. Pan, Z. Chen, and L. Xing, "Developing a dynamic speed control system for mixed traffic flow to reduce collision risks near freeway bottlenecks," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 11, pp. 12560-12581, Nov. 2023.

[14] D. Liu, B. Besselink, S. Baldi, W. Yu, and H. L. Trentelman, "A reachability approach to disturbance and safety propagation in mixed platoons," *IEEE Transactions on Automatic Control*, vol. 69, no. 5, pp. 3206-3213, May 2024.

[15] J. Zhou, L. Yan, and K. Yang, "Enhancing System-Level Safety in Mixed-Autonomy Platoon via Safe Reinforcement Learning," *IEEE Transactions on Intelligent Vehicles*, doi: 10.1109/TIV.2024.3373512.

[16] C. Chen, J. Wang, Q. Xu, J. Wang, and K. Li, "Mixed platoon control of automated and human-driven vehicles at a signalized intersection: dynamical analysis and optimal control," *Transportation Research Part C: Emerging Technologies*, vol. 127, pp. 103138, June 2021.

[17] F. Zheng, C. Liu, X. Liu, S. E. Jabari, and L. Lu, "Analyzing the impact of automated vehicles on uncertainty and stability of the mixed traffic flow," *Transportation Research Part C: Emerging Technologies*, vol. 112, pp. 203-219, Mar. 2020.

[18] J. Yang, D. Zhao, J. Lan, S. Xue, W. Zhao, D. Tian, Q. Zhou, and K. Song, "Eco-driving of general mixed platoons with CAVs and HDVs," *IEEE Transactions on Intelligent Vehicles*, vol. 8, no. 2, pp. 1190-1203, Feb. 2023.

[19] S. Zhu, D. Li, and M. Liu, "Hindrance-aware platoon formation for connected vehicles in mixed traffic," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 12, pp. 24876-24890, Dec. 2022.

[20] S. Jin, D. Sun, and Z. Liu, "The impact of spatial distribution of heterogeneous vehicles on performance of mixed platoon: a cyber-physical perspective," *KSCE Journal of Civil Engineering*, vol. 25, pp. 303-315, Oct. 2020.

[21] K. Li, J. Wang, and Y. Zheng, "Cooperative formation of autonomous vehicles in mixed traffic flow: beyond platooning," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 9, pp. 15951-15966, Sept. 2022.

[22] X. Hu, Z. Zheng, D. Chen, and J. Sun, "Autonomous vehicle's impact on traffic: empirical evidence from Waymo open dataset and implications from modelling," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 6, pp. 6711-6724, June 2023.

[23] S. Woo and A. Skabardonis, "Flow-aware platoon formation of connected automated vehicles in a mixed traffic with human-driven vehicles," *Transportation Research Part C: Emerging Technologies*, vol. 133, pp. 103442, Dec. 2021.

[24] M. Cai, Q. Xu, C. Chen, J. Wang, K. Li, J. Wang, and X. Wu, "Formation control with lane preference for connected and automated vehicles in multi-lane scenarios," *Transportation Research Part C: Emerging Technologies*, vol. 136, pp. 103513, Mar. 2022.

[25] R. Wu, H. Jia, Q. Huang, J. Tian, H. Gao, and G. Wang, "Multi-lane unsignalized intersection cooperation strategy considering platoons formation in a mixed connected automated vehicles and connected human-driven vehicles environment," *IEEE Transactions on Intelligent Transportation Systems*, vol. 25, no. 2, pp. 1569-1585, Feb. 2024.

[26] S. Maiti, S. Winter, L. Kulik, and S. Sarkar, "Ad-Hoc platoon formation and dissolution strategies for multi-lane highways," *Journal of Intelligent Transportation Systems*, vol. 27, no. 2, pp. 161-173, Nov. 2023.

[27] Gopinath, Deepak, J. DeCastro, G. Rosman, E. Sumner, Allison Morgan, Shabnam Hakimi, and Simon Stent, "Highway-env: a framework for simulating behaviors and preferences to support human-AI teaming in driving," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4342-4350, June 2022.

[28] H. Dong, Q. Hu, D. Li, Z. Li, and Z. Song, "Predictive battery thermal and energy management for connected and automated electric vehicles," *IEEE Transactions on Intelligent Transportation Systems*, doi: 10.1109/TITS.2024.3494734.

[29] X. Duan, C. Sun, D. Tian, J. Zhou, and D. Cao, "Cooperative lane-change motion planning for connected and automated vehicle platoons in multi-lane scenarios," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 7, pp. 7073-7091, July 2023.

[30] T. L. Pan, W. H. K. Lam, A. Sumalee, and R. X. Zhong, "Modeling the impacts of mandatory and discretionary lane-changing maneuvers," *Transportation Research Part C: Emerging Technologies*, vol. 68, pp. 403-424, July 2016.

[31] H. Dong, Q. Wang, W. Zhuang, G. Yin, K. Gao, Z. Li, and Z. Song, "Flexible eco-cruising strategy for connected and automated vehicles with efficient driving lane planning and speed optimization," *IEEE Transactions on Transportation Electrification*, vol. 10, no. 1, pp. 1530-1540, Mar. 2024.

[32] H. Dong, W. Zhuang, G. Wu, Z. Li, G. Yin, and Z. Song, "Overtaking-enabled eco-approach control at signalized intersections for connected and automated vehicles," *IEEE Transactions on Intelligent Transportation Systems*, doi: 10.1109/TITS.2023.3328022.

[33] V. Otterlo, Martijn, and M. Wiering, "Reinforcement learning and markov decision processes," *Reinforcement Learning: State-of-the-art*, Berlin, Heidelberg: Springer Berlin Heidelberg, vol. 23, pp. 3-42, 2012.

[34] Mahadevan and Sridhar, "Average reward reinforcement learning: foundations, algorithms, and empirical results," *Machine Learning*, no. 1, vol. 22, pp. 159-195, Jan. 1996.

[35] B. Fang, C. Zheng, H. Wang, and T. Yu, "Two-stream fused fuzzy deep neural network for multiagent learning," *IEEE Transactions on Fuzzy Systems*, vol. 31, no. 2, pp. 511-520, Feb. 2023.

[36] S. Kuutti, R. Bowden, H. Joshi, R. d. Temple, and S. Fallah, "End-to-end reinforcement learning for autonomous longitudinal control using advantage actor critic with temporal context," *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, Auckland, New Zealand, 2019, pp. 2456-2462.

[37] T. Schaul, J. Quan, I. Antonoglou, and D. Silver, "Prioritized experience replay," *arXiv preprint arXiv:1511.05952*, 2015.

[38] H. Alibrahim and S. A. Ludwig, "Hyperparameter optimization: comparing genetic algorithm against grid search and bayesian optimization," *2021 IEEE Congress on Evolutionary Computation (CEC)*, Kraków, Poland, 2021, pp. 1551-1559.

[39] G. Wang, W. Xie, A. Demers, and J. Gehrke, "Asynchronous large-scale graph processing made easy," *CIDR*, vol. 13, pp. 3-6, Jan. 2013.

[40] X. Guo, R. Xu, and T. Zariphopoulou, "Entropy regularization for mean field games with learning," *Mathematics of Operations Research*, vol. 47, no. 4, pp. 3239-3260, Nov 2022.

[41] Z. Yao, Y. Wu, Y. Wang, B. Zhao, and Y. Jiang, "Analysis of the impact of maximum platoon size of CAVs on mixed traffic flow: an analytical and simulation method," *Transportation Research Part C: Emerging Technologies*, vol. 147, pp. 103989, Feb. 2023.

[42] Y. Pan, Y. Wu, L. Xu, C. Xia, and D. L. Olson, "The impacts of connected autonomous vehicles on mixed traffic flow: a comprehensive review," *Physica A: Statistical Mechanics and its Applications*, pp. 129454, Jan. 2024.

[43] Z. H. Khattak, B. L. Smith, M. D. Fontaine, J. Ma, and A. J. Khattak, "Active lane management and control using connected and automated vehicles in a mixed traffic environment," *Transportation Research Part C: Emerging Technologies*, vol. 139, pp. 103648, June 2022.

[44] H. Dong, J. Shi, W. Zhuang, Z. Li, and Z. Song, "Analyzing the impact of mixed vehicle platoon formations on vehicle energy and traffic efficiencies," *Applied Energy*, vol. 377, pp. 124448, Jan. 2025.

[45] Zefreh, M.M. Maghrour, and A. Török, "Distribution of traffic speed in different traffic conditions: an empirical study in Budapest," *Transport*, vol. 35, no. 1, pp. 68-86, Mar. 2020.

[46] D. Chen, L. Jiang, Y. Wang, and Z. Li, "Autonomous driving using safe reinforcement learning by incorporating a regret-based human lane-changing decision model," *2020 American Control Conference (ACC), Denver, CO, USA*, 2020, pp. 4355-4361.

[47] Y. J. Park, Y. J. Lee, and S. B. Kim, "Cooperative multi-agent reinforcement learning with approximate model learning," *IEEE Access*, vol. 8, pp. 125389-125400, 2020.

[48] Han, Jihun, A. Vahidi, and A. Sciarretta, "Fundamentals of energy efficient driving for combustion engine and electric vehicles: an optimal control perspective," *Automatica*, vol. 103, pp. 558-572, May 2019.

[49] A. Coppola, D. G. Lui, A. Petrillo, and S. Santini, "Eco-driving control architecture for platoons of uncertain heterogeneous nonlinear connected autonomous electric vehicles," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 12, pp. 24220-24234, Dec. 2022.

[50] Treiber, Martin, A. Hennecke, and D. Helbing, "Congested traffic states in empirical observations and microscopic simulations," *Physical review*

E, vol. 62, no. 2, pp. 1805, Aug. 2000.

[51] P. Chakroborty, S. Agrawal, and K. Vasishtha, "Microscopic modeling of driver behavior in uninterrupted traffic flow," *Journal of Transportation Engineering*, vol. 130, pp. 438–451, July 2004.

[52] Z. Wu, C. Yu, D. Ye, J. Zhang, and Hankz Hankui Zhuo, "Coordinated proximal policy optimization," *Advances in Neural Information Processing Systems*, vol. 34, pp. 26437-26448, Dec. 2021.

[53] B. He, J. Wang, Q. Qi, H. Sun, J. Liao, C. Du, X. Yang, and Z. Han, "DeepCC: multi-agent deep reinforcement learning congestion control for multi-path TCP based on self-attention," *IEEE Transactions on Network and Service Management*, vol. 18, no. 4, pp. 4770-4788, Dec. 2021.

[54] S. Li, Y. Wu, X. Cui, H. Dong, F. Fang, and Russell S, "Robust multi-agent reinforcement learning via minimax deep deterministic policy gradient.," *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, pp. 4213-4220, July 2019.